openGauss AI特性创新实践课



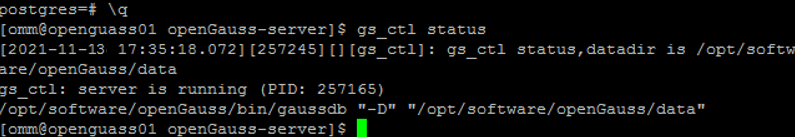
华为技术有限公司

# 关卡一、openGauss数据安装及基本操作

openGauss数据安装及基本操作, 作业提交任务如下：

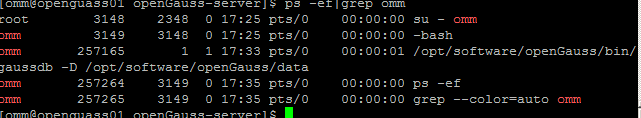
任务一：数据库状态验证

1. 查询数据库状态成功截图



任务二：数据库服务进程验证

1. 查看数据库服务进程截图（包含数据库服务器的主机名）



实验思考题：为什么需要通过源码编译，安装数据库？

源代码可以允许用户选择因为软件在编译期间需要配置，不同用户。openGauss 数据库通过编译安装，是因为要定制存储引擎。此外，软件也需要统一的linux内核版本可能不同，计算机安装路径。

安装数据库所需要的硬件也不同，编译器步骤：

1. 安装好第三方编译库binarylibs和编译管理器cmake
2. 下载openGauss数据库的源码
3. 安装对应依赖包
4. 设置用户的环境变量且使生效
5. 进入Openguass源码下生成配置文件
6. 使用make+cmake编译且安装数据库

# 关卡二、openGauss数据导入及基本操作

任务一：数据初始化验证

1. 查询supplier表的行数，并将结果进行图：

select count(\*) from supplier;;

Text

Description automatically generated

任务二：行存表与列存表执行效率对比

1. 2020年上半年litemall\_orders行存表与litemall\_orders\_col列存表中的order\_price的总和查询，并对比执行效率截图

select sum (order\_price) from litemall\_orders where add\_date between '20200101' and '20200701';

Graphical user interface, text

Description automatically generated

select sum (order\_price) from litemall\_orders\_col where add\_date between '20200101' and '20200701';

Text

Description automatically generated

2. 2020年上半年litemall\_orders行存表与litemall\_orders\_col列存表中的order\_price的平均值查询，并对比执行效率截图

select avg (order\_price) from litemall\_orders where add\_date between '20200101' and '20200701';

Text

Description automatically generated

select avg (order\_price) from litemall\_orders\_col where add\_date between '20200101' and '20200701';

Text

Description automatically generated

3. 查询litemall\_orders行存表与litemall\_orders\_col列存表中order\_id为6的order\_price的值，并对比执行效率截图。

select order\_price from litemall\_orders where order\_id=6;

Text

Description automatically generated

select order\_price from litemall\_orders\_col where order\_id=6;

Text

Description automatically generated

4. 将litemall\_orders行存表与litemall\_orders\_col列存表中order\_id为6的order\_price修改为2468，并对比执行效率截图。

update litemall\_orders set order\_price=2468 where order\_id=6;

Text

Description automatically generated

update litemall\_orders\_col set order\_price=2468 where order\_id=6;

Text

Description automatically generated with medium confidence

任务三：物化视图的使用

1. 创建物化视图所需要的表后，对表内容进行查询，对查询结果截图：

SELECT \* FROM test\_view;

A picture containing text

Description automatically generated

2. 使用物化视图统计人数，查询物化视图结果，将执行结果截图。

SELECT \* FROM v\_order;

Text

Description automatically generated

3. 对表进行操作后，刷新物化视图，查询物化视图结果，将执行结果截图。

SELECT \* FROM v\_order;

Text

Description automatically generated

4. 创建增量物化视图，查询物化视图结果，将执行结果截图。

SELECT \* FROM vi\_order;

A picture containing text

Description automatically generated A picture containing table

Description automatically generated

5. 对表进行操作后，刷新增量物化视图，查询物化视图结果，将执行结果截图。

SELECT \* FROM vi\_order;

A picture containing table

Description automatically generated

实践思考题1：行存表与列存表在执行相同的SQL语句时，为何执行的时间不同？在执行哪些类型SQL时，行存表效率更高？在执行哪些类型SQL时，列存表效率更高？

1-2.行存表与列比行快约11倍

3-4.列比行快约23倍

4-5.行比列快约3倍

7-8.行比列快约20倍

在1-4图中，存在年份date数据的判定和价钱numeric数据的相加或平均值，行存储需要读完所有数据，存在冗余列。而列存储只需要读取两列数据，不存在冗余性问题。并且列存储的每一列数据类型是同质的，不存在二义性问题。比如说价格数据类型为浮点数(numeric)，那么它的数据集合一定是浮点数数据。这种情况使数据解析变得十分容易。相比之下，行存储则要复杂得多，因为在一行记录中保存了多种类型的数据，数据解析需要在多种数据类型之间频繁转换，这个操作很消耗CPU，增加了解析的时间。所以，此时列存储的解析过程更有利于分析数据。

在5-8图中

行存储只需要读取每行的order\_id，相当于读一列完成；

列存储需要读取order\_id列，再读取order\_price列，在读取上花时就比行存储用时多，再加上列存储的修改和添加数据意味着磁头调度次数多，而磁头调度是需要时间的，一般在1ms~10ms，再加上磁头需要在盘片上移动和定位花费的时间，实际时间消耗会更大。

总而言之，如果关注整张表的内容，而不是单独某几列，并且所关注的内容是不需要通过任何聚集运算的，那么推荐使用行式存储，效率更高。反之，关注的都是某几列的内容，或者有频繁聚集需要的，通过聚集之后进行数据分析的表那么用列存储效率更高。

数据读取存表在执行相同的SQL语句时，执行时间不同。因为行存储通常将一行数据完全读出是一次性完成的，而列存储在每次读取的数据是集合中的一段或者全部。

数据读取存表在执行相同的SQL语句时，执行时间不同。因为行存储通常将一行数据完全读出是一次性完成的，而列存储在每次读取的数据是集合中的一段或者全部。

在执行数据查询和修改数据类型的SQL时，行存表效率更高。

在执行总和查询和平均值查询的SQL时，列存表效率更高。

实践思考题2：全量物化视图与增量物化视图有哪些差别？

全量物化视图只能支持对于创建好的区划试图做全量更新，但是不能支持做增量更新。创建全量物化视图语法和CREATE TABLE AS 语法不一样，不能支持对全量物化指定的NodeGroup创建

增量物化视图顾名思义就是可以对物化视图增量刷新，需要用户手动执行语句完成对物化视图在一段时间内的增量数据进行刷新。

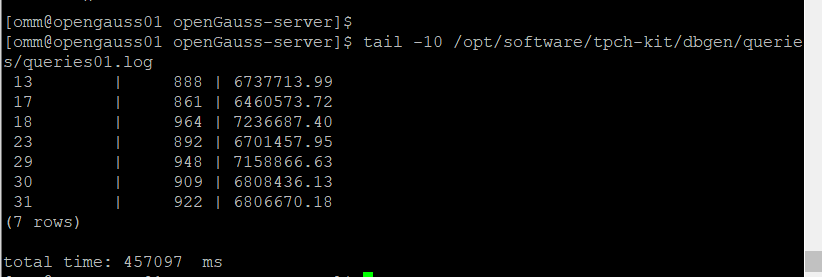
这两者不同的地方就是增量物化视图所支持的场景较小，物化视图创建的语句也只是支持及表扫描语句或者UNION ALL语句。

# 关卡三、openGauss的AI4DB特性应用

任务一：使用X-Tuner进行参数优化

1. 执行TPCH脚本，获得测试时间，将执行结果截图：

gsql -d tpch -p 5432 -r -f /opt/software/tpch-kit/dbgen/queries/queries.sql > /opt/software/tpch-kit/dbgen/queries/queries01.log



2. 使用root用户，执行X-Tuner进行参数建议优化，将执行结果截图

gs\_xtuner recommend --db-name tpch --db-user omm --port 5432 --host 127.0.0.1 --host-user omm

Text

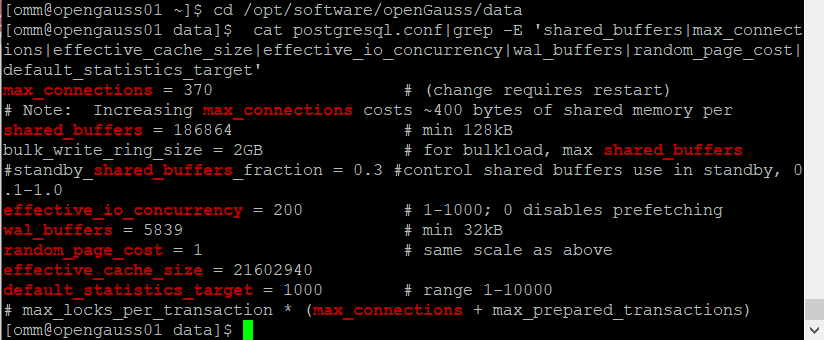
Description automatically generatedA picture containing graphical user interface

Description automatically generated

3.重启完成后，获取参数值：

cd /opt/software/openGauss/data

cat postgresql.conf|grep -E 'shared\_buffers|max\_connections|effective\_cache\_size|effective\_io\_concurrency|wal\_buffers|random\_page\_cost|default\_statistics\_target'



任务二：使用Index-advisor对select 查询语句进行优化，并通过对比执行计划，得到优化前后的不同。

1. 使用explain，对查询2020年3月订单表收入并进行排序的SQL加以分析，将结果截图。

EXPLAIN

SELECT ad.province AS province, SUM(o.actual\_price) AS GMV

FROM litemall\_orders o,

address\_dimension ad,

date\_dimension dd

WHERE o.address\_key = ad.address\_key

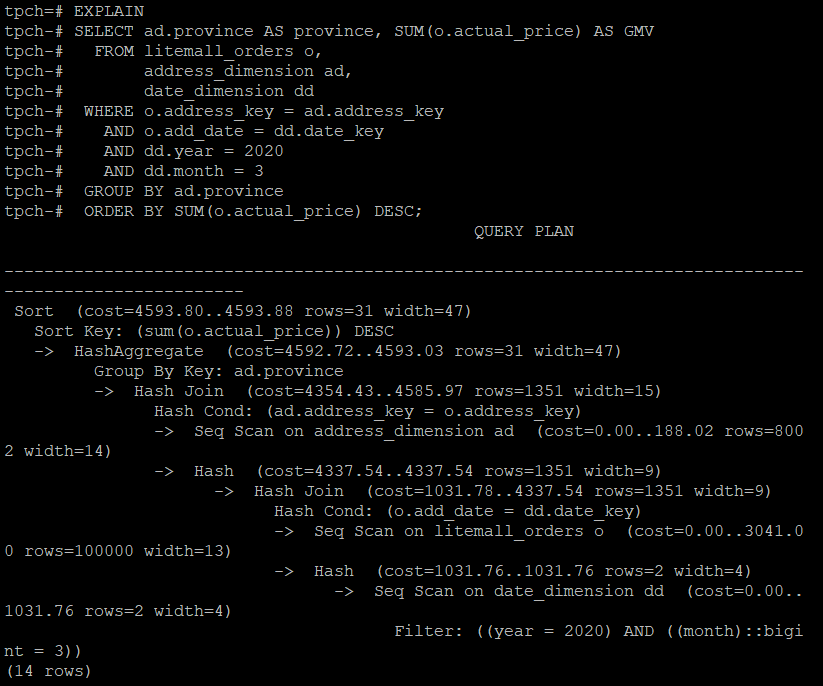
AND o.add\_date = dd.date\_key

AND dd.year = 2020

AND dd.month = 3

GROUP BY ad.province

ORDER BY SUM(o.actual\_price) DESC;

Text

Description automatically generated

2. 使用索引推荐功能，对查询语句进行推荐，将执行结果截图。

select \* from gs\_index\_advise('

SELECT ad.province AS province, SUM(o.actual\_price) AS GMV

FROM litemall\_orders o,

address\_dimension ad,

date\_dimension dd

WHERE o.address\_key = ad.address\_key

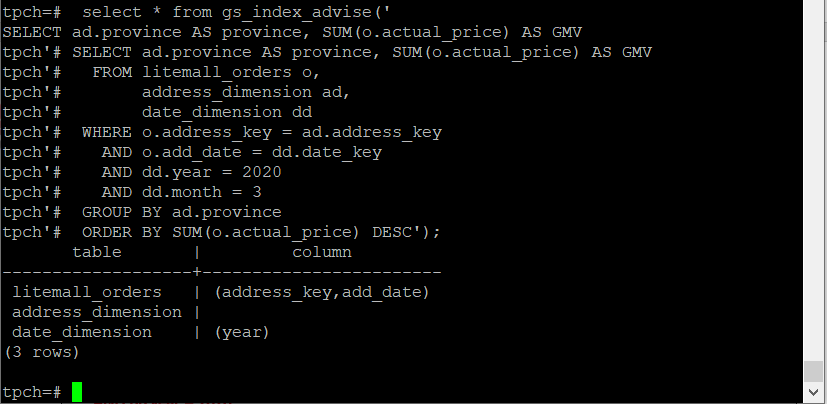
AND o.add\_date = dd.date\_key

AND dd.year = 2020

AND dd.month = 3

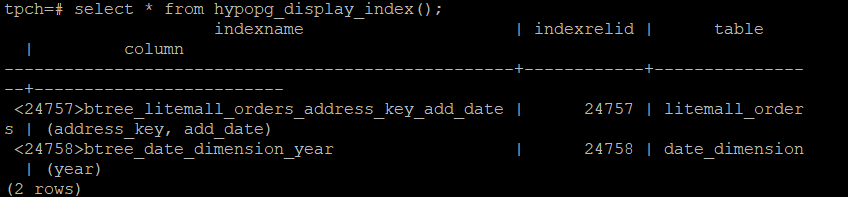
GROUP BY ad.province

ORDER BY SUM(o.actual\_price) DESC');



3. 查看创建的虚拟索引列，将执行结果截图。

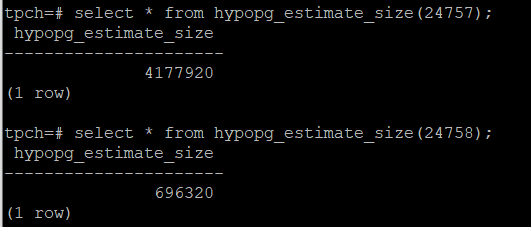
select \* from hypopg\_display\_index();



4. 获取索引虚拟列大小结果（单位为：字节），将执行结果截图。

select \* from hypopg\_estimate\_size(24757);

select \* from hypopg\_estimate\_size(24758);



5.再次使用explain，对该SQL加以分析，将执行结果截图。

EXPLAIN

SELECT ad.province AS province, SUM(o.actual\_price) AS GMV

FROM litemall\_orders o,

address\_dimension ad,

date\_dimension dd

WHERE o.address\_key = ad.address\_key

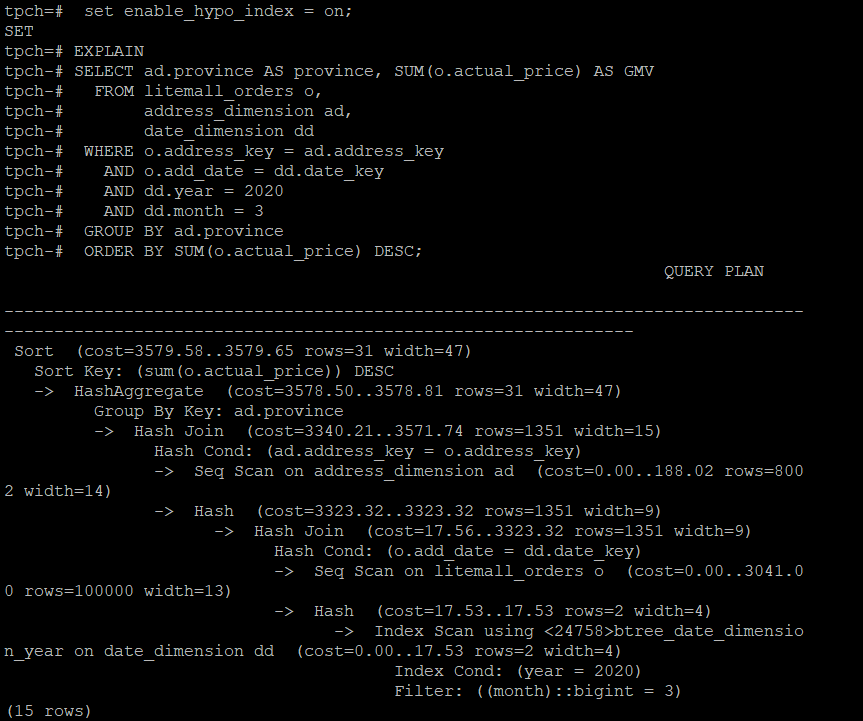
AND o.add\_date = dd.date\_key

AND dd.year = 2020

AND dd.month = 3

GROUP BY ad.province

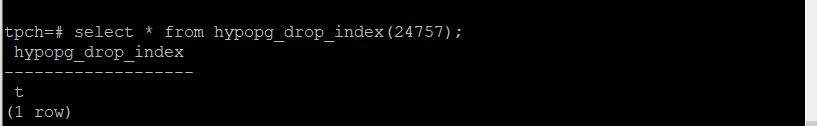
ORDER BY SUM(o.actual\_price) DESC;

Text

Description automatically generated

6. 删除某一个索引虚拟列，将执行结果截图。

select \* from hypopg\_drop\_index(24757);



7. 删除某一个索引虚拟列，将执行结果截图。

select \* from hypopg\_reset\_index();

Text

Description automatically generated

8. 查看索引虚拟列，将执行结果截图。

select \* from hypopg\_display\_index();

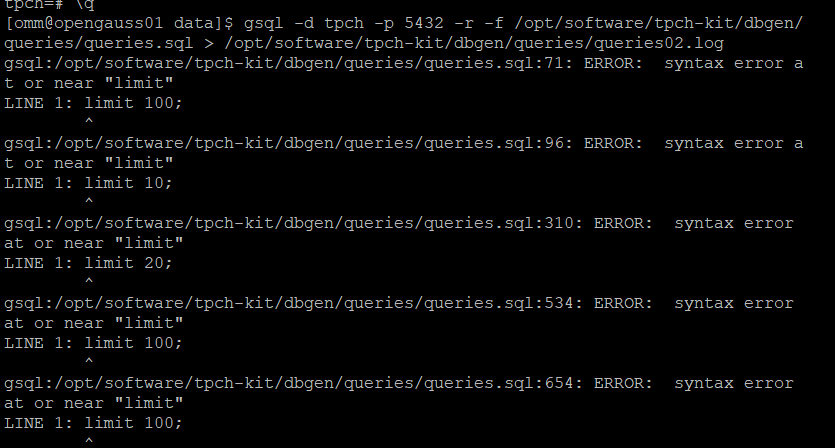
A picture containing graphical user interface

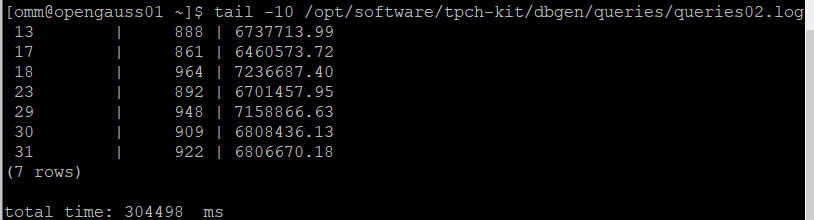
Description automatically generated

任务三：通过创建索引，对queries.sql中的SQL语句进行优化，并对比优化前后queries.sql执行的时间。

1. 重新执行queries.sql查询，将执行结果截图：

gsql -d tpch -p 5432 -r -f /opt/software/tpch-kit/dbgen/queries/queries.sql > /opt/software/tpch-kit/dbgen/queries/queries02.log





挑战一：进一步优化queries.sql中的查询语句，使得前后执行时间出现倍数级的提升。

1. 重新执行queries.sql查询，将执行结果截图：

gsql -d tpch -p 5432 -r -f /opt/software/tpch-kit/dbgen/queries/queries.sql > /opt/software/tpch-kit/dbgen/queries/queries03.log

<https://opengauss.org/zh/docs/2.0.0/docs/Developerguide/SMP%E4%BD%BF%E7%94%A8%E5%BB%BA%E8%AE%AE.html>并行查询

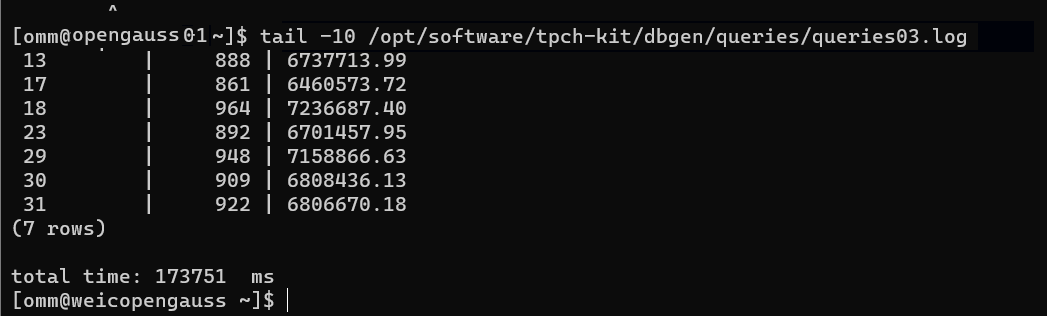
示例代码,要在.sql里每个查询和子查询前都加上SET query\_dop = 4;

openGauss=# SET query\_dop = 4;

openGauss=# SELECT COUNT(\*) FROM t1 GROUP BY a;

......

openGauss=# SET query\_dop = 1;



实践思考题1：根据X-Tuner给出的参数优化，在哪些参数上进行了优化，为何要对这些参数进行优化？

2个参数，max\_connections，shared\_buffers

max\_connections

max\_connections 决定了客户端的最大并发连接数，默认值通常为 100。如果出现连接数过多，无法连接数据库的错误时，可能需要考虑增加最大连接数。不过，修改该参数还需要考虑对其他参数的影响（尤其是 work\_mem）；因为它们是基于每个连接设置的值，增加连接数也会导致这些内存使用量的增加。

shared\_buffers

除了操作系统的 I/O 缓存之外，PostgreSQL 还会使用自己的内部缓存。PostgreSQL 共享内存缓冲区由参数 shared\_buffers 设置，它决定了 PostgreSQL 能够使用的专用缓存大小。

为确保在所有机器和操作系统上的兼容性，PostgreSQL 默认将该值设置得很小，通常是 128 MB。因此，增加 shared\_buffers 的值是提高性能最有效的设置之一。

修改 shared\_buffers 之后必须重启服务器才能生效

A picture containing graphical user interface

Description automatically generated

实践思考题2：索引的使用，对于执行SQL有什么好处？除了使用索引和参数外，还有哪些方面可以对数据库进行优化？

数据库索引是为了增加查询速度而对表字段附加的一种标识。如果我们对某一字段增加索引，查询时就会先去索引列表中一次定位到特定值的行数，大大减少遍历匹配的行数，所以能明显增加查询的速度。

除了使用索引和参数外，还可以通过SQL语句以及有效索引、数据结构、系统配置和硬件来对数据库进行优化。

1. 设计表时选好字段类型，对数据量较大的表，为减少与其他表的关联，增加一些冗余字段提高查询性能。
2. 根据需求要写结构良好的SQL，再根据SQL在表中建立有效的索引
3. 硬件配置的优化。由于openGauss支持使用SSD盘作为数据的组存储设备，但必须是SAS接口的SSD盘，以RAID的方式部署使用。服务器不能太差，磁盘的读写性能影响io
4. 对访问频繁的数据，充分利用数据库cache和应用的缓存
5. 数据库实体的碎片的整理.包括使用分区表.

# 关卡四、openGauss的DB4AI特性应用

任务一：在gs\_model\_warehouse系统表中查看训练后的模型信息，将执行结果截图：

postgres=# SELECT \* FROM gs\_model\_warehouse WHERE modelname = 'house\_binary\_classifier';

A picture containing text, screenshot, monitor, black

Description automatically generated

任务二：观察新模型的信息，将执行结果截图。

postgres=# SELECT \* FROM gs\_model\_warehouse WHERE modelname = 'house\_binary\_classifier';

A screenshot of a computer

Description automatically generated

任务三：利用训练好的逻辑回归模型预测数据，并与SVM算法进行比较，将执行结果截图。

postgres=# SELECT tax, bath, size, price, price < 100000 AS price\_actual, PREDICT BY house\_binary\_classifier (FEATURES tax, bath, size) AS price\_svm\_pred, PREDICT BY house\_logistic\_classifier (FEATURES tax, bath, size) AS price\_logistic\_pred FROM houses;

Graphical user interface, text

Description automatically generated

实践思考题1：分类模型与回归模型有何不同？

分类模型与回归模型的区别主要在于输出变量类型是否连续，预测连续值的为回归问题、预测离散值的为分类问题。回归模型输出变量的类型也成为定量输出，或者连续变量预测；而分类模型则成为定性输出，或者离散变量预测。分类模型可以将回归模型的输出离散化，而回归模型可以将分类模型的输出连续化。

实践思考题2：什么是SVM算法？

支持向量机(support vector machine)是一种分类SVM算法，通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。通俗来讲，它是一种二类分类模型，其他的基本模型可以定义为特征空间上的间隔最大的线性分类器，即支持向量机的其学习策略便是间隔最大化，最终可最后可以转化为一个凸二次规划问题的求解。

实践思考题3：分类问题有哪些评价指标，请分别说明他们的含义？

将实力分成正类 (positive) 或负类 (negative)，在实际分类中会出现一下四种情况：

1. 若一个实力是正类，并且被预测为正类，即是真正类TP (True Positive)
2. 若一个实力是正类，并且被预测为负类，即是假负类FN (False Negative)
3. 若一个实力是负类，并且被预测为正类，即是假正类FP (False Positive)
4. 若一个实力是负类，并且被预测为负类，即是真负类TN (True Negative)

* 准确率：(Accuracy-所有正确预测文档数与总数的比率TP+TN/TP+FP+FN+TN)

准确率的缺点是数据的样本不平均，不能用来评价模型的性能优劣

* 精确率：(Precision-检索出相关文档数与检索出的文档总数的比率TP/TP+FP)

精确率是在模型预测为正样本的结果中，真正是正样本所占的百分比，分类阈值比较高。

* 召回率：(Recall-所有准确的条目的比例TP/TP+FN)

在史记为正样本中，被预测为正样本所占的百分比，它尽量检测数据，不遗漏数据，所以分类阈值较低。

* F1 Score

精准率和召回率都有各自的缺点，如果阈值较高，精准率就会高，但同时召回很多数据，反之阈值较低，召回率就会高，但预测的不准确。

F1 Score 是一种调和平均数。

* P-R曲线(Precesion-Recall Curve)

对于所有正样本，用来描述精准率和召回率变化的曲线。

模型月座标围城的面积越大，模型的性能就越好。由于曲线下的面积很难估算，BEP (Break-Event Point)，也称为平衡点，就是当P = R时的取值，平衡点的取值越高，性能更优。

* ROC (Receiver Operating Characteristic)

又称为接收者操作特征曲线。理想的目标为TPR=1, FPR=0。所以ROC曲线越靠拢(0,1)点，越偏离45度对角线越好，对应的就是TPR越大越好，FPR越小越好。

* AUC (Area Under Curve)

这个曲线是处于ROC曲线下的那部分面积的大小。AUC越大，代表模型的性能越好。

实践思考题4：回归问题有哪些评价指标，请分别说明他们的含义？

* 平均绝对误差 (MAE, Mean Absolute Error)对绝对误差损失的预期值,绝对误差的平均值对应于平方（二次）误差的期望
* 平均绝对百分比误差 (MAPE, Mean Absolute Percentage Error)对相对误差损失的预期值，也就是绝对误差和真值的百分比
* 均方误差 (MSE, Mean Squared Error)对应于平方（二次）误差的期望
* 均方误差跟 / 均方根误差(RMSE, Root Mean Squared Error)衡量观测值与真实值之间的偏差
* 均方误差对数 (MSLE, Mean Squared Log Error)对应平方对数（二次）差的预期
* 中位绝对误差 (MedAE, Median Absolute Error)通过取目标和预测之间的所有绝对差值的中值来计算损失
* R Squared (R2 Score)反映了自变量x对因变量y的变动的解释的程度。越接近1，说明模型的拟合越好。

表示了y的变动的程度，正比于方差

表示了模型和真实值的残差

表示了模型对y的变动的预测